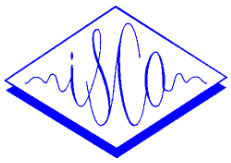

Second ESCA/IEEE Workshop
on Speech Synthesis
Mohonk, New Paltz, NY, USA
September 12-15, 1994

Building prototypes for articulatory speech synthesis

Gérard Bailly[1], Eric Castelli[1] and Bernard Gabioud[2]

[1] Institut de la Communication Parlée - INPG/Université Stendhal - Grenoble - France
[2] Laboratoire d'Analyse Informatique de la Parole - University of Lausanne - Switzerland

Abstract

We present a method for specifying acoustic targets for articulatory synthesis. The controller is driven by an "acoustic" score which specifies the skeleton of the desired audio-visual properties of the output. We examine here the possibility of specifying simple VV and CV transitions.

Introduction

Basic terminal-analogue speech synthesisers used in current rule synthesis have sufficient degrees of freedom to perform a close-copy of various natural voice types. Such flexible and dextrous synthesisers are highly desirable to improve the "naturalness" of synthetic speech. Keeping this flexibility has always been claimed in the literature [10] in order to compensate for the poorer speech quality obtained compared to concatenative speech synthesis. Such a flexibility is however an enormous drawback when the output degrees of freedom are far away from the real degrees of freedom of the human speaker: for example, the formant space is of course not an hyper cube (see Fig. 2). Sophisticated control models [5, 11] have to be implemented to palliate the lack of speech specificity of the "instrument" and thus avoid the generation of acoustic "monsters".

We claim that the most appropriate solution for implementing production (producing ecological sounds) and perception (sounds belonging to a particular phonological system) constraints is a judicious use of articulatory synthesis, aerodynamic and acoustic simulation of the airflow through the shaped cavities and motor control principles. Following the opposite but complementary trail to Stevens and Bickley, we propose to control articulatory synthesis by acoustic constraints, whereas HL parameters, although incorporating acoustic (formants, fundamental frequency), aerodynamic (intraoral pressure) and geometric (constrictions) cues, are claimed at being "quasi-articulatory".

1. The articulatory model

We use an improved version of the Maeda's model [3] based on a re-analysis of the 519 tracings of mid-sagittal radiographs used by the original statistical analysis. A better model of the position of the apex was obtained by applying a linear regression analysis on the residual of a factorial analysis applied on the medial tongue region excluding the apex and the larynx region: the deviation of the tongue tip is then described with a resolution of ±1 mm (see Fig. 1) instead of ±2.3 in the original model. The seven factors of the successive analysis have the same interpretation as those of Maeda but guaranty a better precision in the front region of vocal tract essential to a proper control of the front cavity resonance.

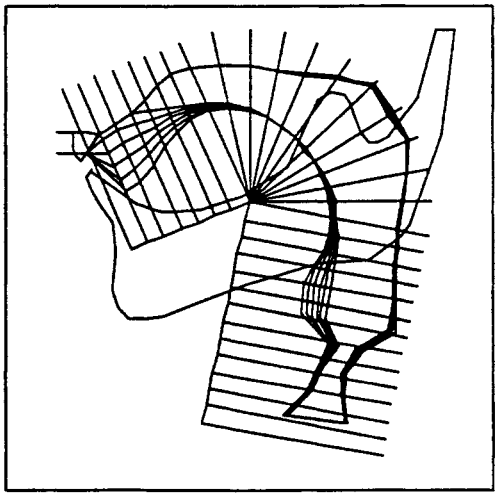

Fig.1: The articulotron: moving the apex from the neutral position. A jaw has been added to the drawing to show its intrinsic movement.

2. The acoustic model

The model built at I.C.P. (named **SIMOND** for **SIM**ulation **O**rale-**N**asale **D**ynamique) is divided into three main parts: an analogue model of the lungs and trachea, a mechanical model of the vocal folds, and an analogue model of the vocal tract. The two analogue parts are acoustically based on the Kelly & Lochbaum principle in [7]. Dynamic equations for the propagation of the sound waves the vocal tract are used in order to simulate fast dynamic variations of the vocal tract geometry [8]. The lungs and the trachea are considered, with the first approximation of a perfect symmetry, as a single tube. Dimensions of the subglottic system (areas and lengths) are consistent with the literature. The trachea has a constant geometry during simulations, only the respiratory zone of the lungs, have variable area. In order to reproduce

the numerous pulmo-alveolar terminations, infinite losses are considered as a model which produces a realistic glottal air-flow. The command parameters are : Dp, the pressure drop across the glottis; Q, the mass-tension factor and Ag0, the opening area of the glottis at rest. The coupling between the mechanical two-mass model and the two analogue parts has been improved [13]: the vocal source is considered as a glottal airflow source applied at an elementary tube which presents the same acoustic characteristics than the other tubes of the analogue parts. The area of this "glottal" tube is equal to the instantaneous area of the glottis, computed with the mechanical two-mass model. The acoustic consequences of this coupling method, especially on the frequencies and bandwidths of the first two formants, had been systematically validated for vowel and for consonant configurations.

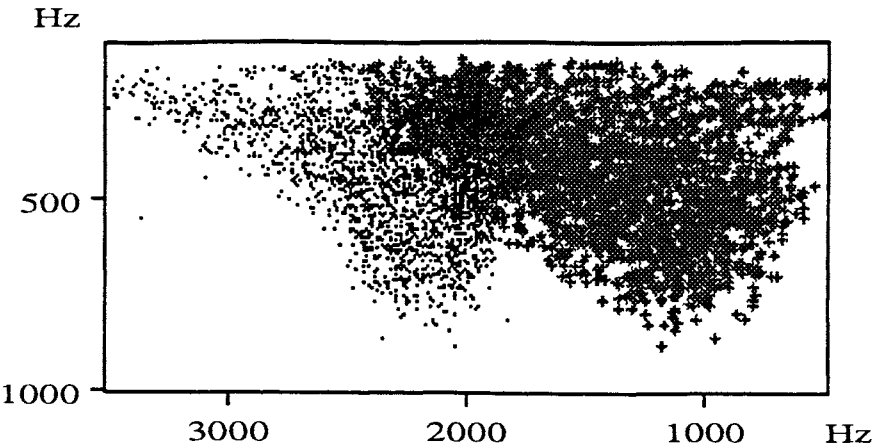

Fig.2: Acoustic space of the articulotron showing how spaces F_1-F_2 and F_1-F_3 overlap. Note how the vocalic triangle is defined.

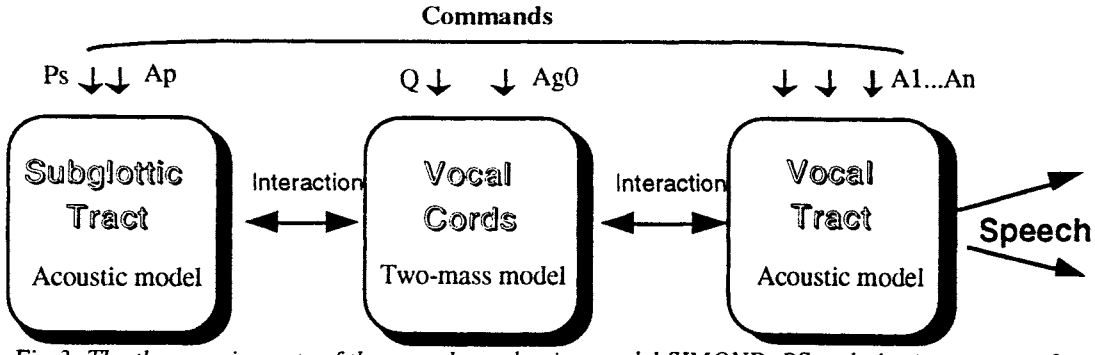

Fig.3: The three main parts of the speech production model SIMOND. PS: subglottic pressure, Q and Ag0 for the vocal cords, A_p, A_1.. A_n: area function.

Losses in the analogue parts of the vocal tract were implemented. Viscosity and heat losses depend upon the area of the elementary tube and of the frequency [8]. Wall vibrations are simulated using a tube coupled in parallel of each elementary tube of the vocal tract and terminated by an impedance. Radiations at the lips or at the nostrils are approximated by a piston in an infinite plane. The nasal tract geometry is quite complex and other cavities, like sinuses, are coupled with it. In order to simulate nasal sounds, our acoustic model is able to calculate sound waves propagation in a complex tract formed by the oral tract, the nasal one and sinuses cavities [4]. In order to simulate continuous length variations of the vocal tract, sound wave propagations are calculated with variable elementary tube lengths, thus, with variable sample frequency. A signal processing procedure produces the output radiated pressure signal at a fixed frequency. An automatic noise source is implemented in the vocal tract for the generation of fricative consonants and the fricative noise is injected at the place of the constriction which is automatically determined.

3. The control model

The general framework for building a control model for an articulatory plant, the *articultron* with excess degrees of freedom has been previously described in [1].The controller is build on a forward model which captures the articulatory-to-acoustic transform. The controller can then use this analytic model to predict acoustic consequences of his gestures and thus tune

articulatory movements to given acoustic targets thanks to back-propagation: aeroacoustic and articulatory improvements of the plant can then permanently benefit to the controller and simplify the control strategies.

The biological plausibility of the articulatory and acoustic models described above is essential to the control model: such features as the coupling between the vocal tract and the nasal and subglottic tracts explains already complex phenomenon such as pole-zeros

3.1. Building vocalic prototypes for the ten French vowels

Once the maximal space of the articulotron has been built using a "babbler" (see Fig.2), we placed acoustic target intervals for the ten French vowels /a, ɛ, e, i, y, ø, œ, ɔ, o, u/. Starting from an initial neutral configuration (all articulatory parameters equal to 0), the gradient descent converge towards prototypic articulatory configurations which will be used to generate the following articulatory trajectories.

3.2. Acoustic vs. articulatory control of VV trajectories

Acoustic measurements of natural VV transitions and results of mimicking experiments [2] show that despite the possibility to join vocalic targets via linear trajectories in the formant space, speakers do not actually act accordingly: the kinematics of acoustic trajectories seems to signal targets by direct paths in the space of the first resonances of the front (R_2) and back cavities (R_1), whereas linear articulatory interpolations fail to explain the observed data.

3.3. Prototypes for occlusives

Acoustic specification of occlusives has been intensively studied in the literature [12]. The theory of relative invariance postulates that the formant loci of occlusives is specified relative to the underlying vocalic gesture. We specified the sensory-motor targets for occlusives with an acoustic gauge following the locus equations given the authors mentioned above. The specification of only three formants is able to explain most of the perception errors ([di] vs. [gi]) but prevent the inversion procedure to converge towards systematic articulatory strategies. We compensate for the insufficient acoustic description of certain targets by imposing a systematic recruitment of specific articulators for each occlusive, i.e. lip and jaw raising for [b], tongue tip raising and dorsum lowering for [d], and tongue tip lowering and dorsum raising for [g]. The computed locus equations are given Fig.5. Our simulations confirm that the popular F_2-F_3 convergence observed for [g] is explained by an exchange of affiliation between formants at around 2000 Hz. When expressed in terms of relative invariance of change in the first front cavity resonance, locus equations are highly linear even for [g] (see the good alignement of $F_{2g}=f(F_{2v})$ for low vowels and $F_{3g}=f(F_{3v})$ for high vowels in Fig.5).Such a linearity is highly desirable in the task space since it describes an articulatory-to-acoustic transform which is easier to learn and inverse.

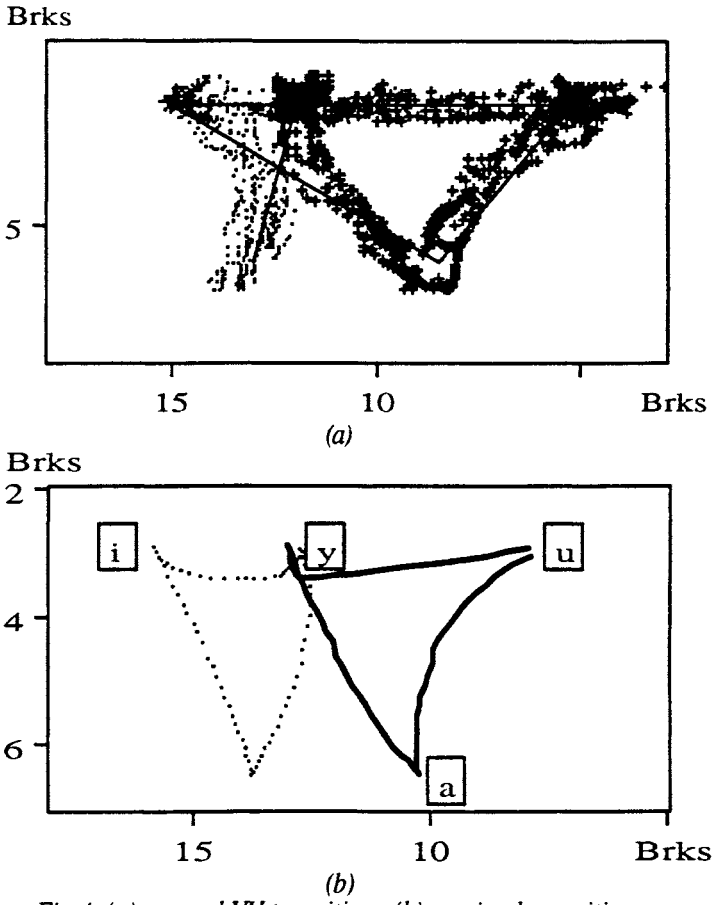

Fig.4. (a) natural VV transitions (b) maximal transitions obtained by a linear interpolation of articulatory configurations.

Conclusions

We demonstrate how articulatory prototypes for vowels and then voiced occlusives can be

build using acoustic gauges. We consider that proprioceptive information such as constrictions, pressures... can then emerge from this negotiation between acoustic requirements and articulatory constraints. This information will then be used to drive and accelerate the convergence of the acoustic-to-articulatory inversion and implement anticipatory behaviour: in our framework, anticipatory behaviour is due to starting positions of the articulators which foreshadow incoming targets. The "articulotron" thus become more and more skilled as it get experience by clustering the solutions of previous inversions.

Our prototypes are virtual targets. These targets are then encoded as equilibrium points [9] for each articulator. The kinematics of the trajectories will be then greatly influenced by the state of the musculo-skeletal system which is in charge with the execution of the movement: the global stiffness of the movement is tuned in parallel to its virtual specification according to prosodic requirements (acoustic enhancement due to hyper-articulation).

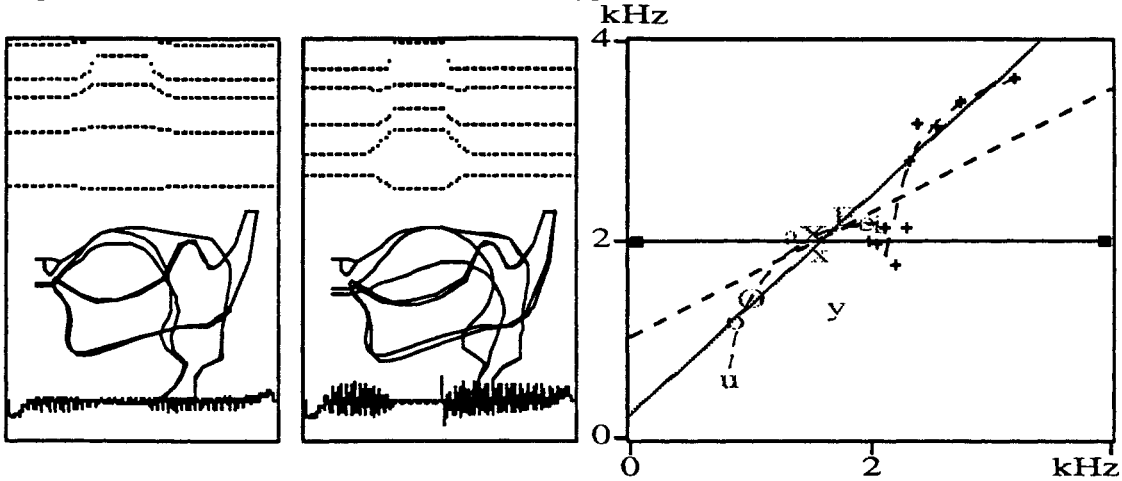

Fig.5: From left to right: the articulatory and acoustic movements for [i] and [a], locus equa-tions for [g] (Sussman et al's vs. R_2 and the almost constant R_3 loci are drawn resp. with dotted and solid lines. $F_{2g}=f(F_{2v})$ and $F_{3g}=f(F_{3v})$ are shown with symbols and crosses)

References

[1] Bailly G., Laboissière R. & Schwartz J.L. (1991) "Formant trajectories as audible gsetures: an alternative for speech synthesis", *Journal of Phonetics*, 19-1, 9-23.

[2] Bailly G. (to appear) "Caracterisation of formant trajectories by tracking vocal tract resonances", in *Levels in Speech Communication: relations and interactions* (J. Schoentgen, J.-M. Ramlot, C.Sorin, H. Méloni and J. Mariani, Eds).

[3] Beautemps D. & Gabioud B. (1994) "Adaptation d'un modèle articulatoire à un locuteur, dans le but de contraindre l'inversion articulatori-acoustique", *XXe Journées d'Etudes sur la Parole*, 119-124.

[4] Castelli E. & Feng G. (submitted) "Some acoustic features of nasality. A new concept of nasality target", *J. Acoust. Soc. Am.*

[5] Coleman J. (1992) "Synthesis-by-rule without segments or rewrite rules", in *Talking Machines: theories, models and applications* (G. Bailly and C. Benoît, Eds), 43—60.

[6] Ishizaka K. & Flanagan J.L. (1972) "Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Cords", *B.S.T.J.*, 51, 1233-1268.

[7] Kelly J.L. & Lochbaum C.C. (1962) "Speech Synthesis", 4th Int. Congr. Acoust., G42.

[8] Liljencrants J. (1985) "Speech Synthesis with a Reflection-Type Line Analog". Thesis Dr. Sciences, R.I.T. Stockholm.

[9] Loevenbruck H. and Perrier P. (1993) "Vocalic reduction: prediction of acoustic and articulatory variabilities with invariant motor commands", *EuroSpeech'93*, 85—88.

[10] Pols L. C. W. and van Bezooijen R. (1991) "Gaining phonetic knowledge whilst improving synthetic speech quality", *Journal of Phonetics*, 19, 139—146.

[11] Stevens K. N. and Bickley C. A. (1991) "Constraints among parameters simplify control of Klatt formant synthesizer", *Journal of Phonetics*, 19, 161—174.

[12] Sussman H.M., McCaffrey H.A. & Matthews S.A. (1991) "An investigation of locus equations as a source of relational invariance for stop place categorization", *J. Acoust. Soc. Am.*, 90.3, 1309-1325.

[13] Trinh Van L., Guérin B. & Castelli E. (1991) "Source-Tract Coupling and the Subglottal System in an Articulatory Synthesizer", *Eurospeech'91*, Genova, 1, 267-270.